LEAD SCORE CASE STUDY

BY:- ADARSH DALMIA

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

SOLUTION METHODOLOGY

Data cleaning and data manipulation.

- Check and handle duplicate data.
- Check and handle NA values and missing values.
- Drop columns, if it contains large amount of missing values and not useful for the analysis.
- Imputation of the values, with median for numerical column and with mode for categorical column.
- Check and handle outliers in data.

EDA (Exploratory Data Analysis)

- **Univariate Analysis**: Assess the distribution and frequency of individual variables using value counts and histograms.
- **Bivariate Analysis**: Explore relationships between two variables with correlation coefficients and scatter plots.
- **Multivariate Analysis**: Examine interactions among multiple variables using pair plots, heatmaps, and PCA.

Hypothesis Testing

Feature Scaling & Dummy Variables and encoding of the data.

Classification technique: logistic regression used for the model making and prediction.

Validation of the model.

Model presentation.

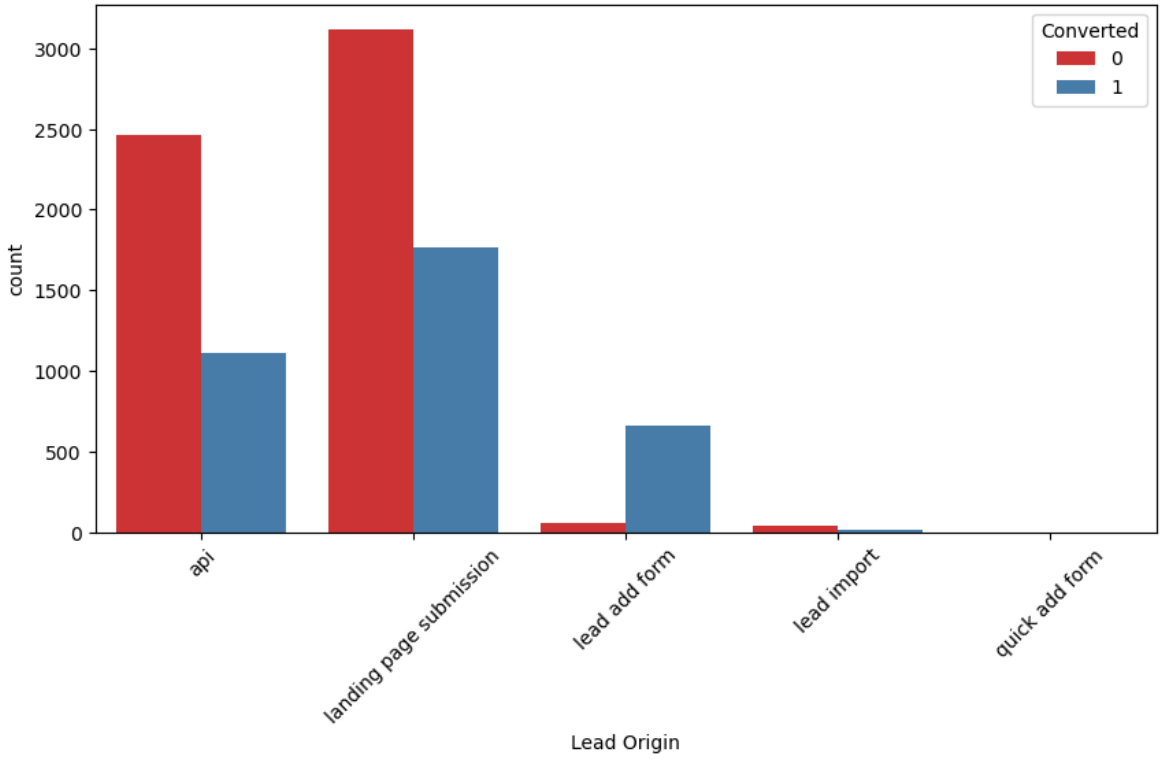
Conclusions and recommendations.

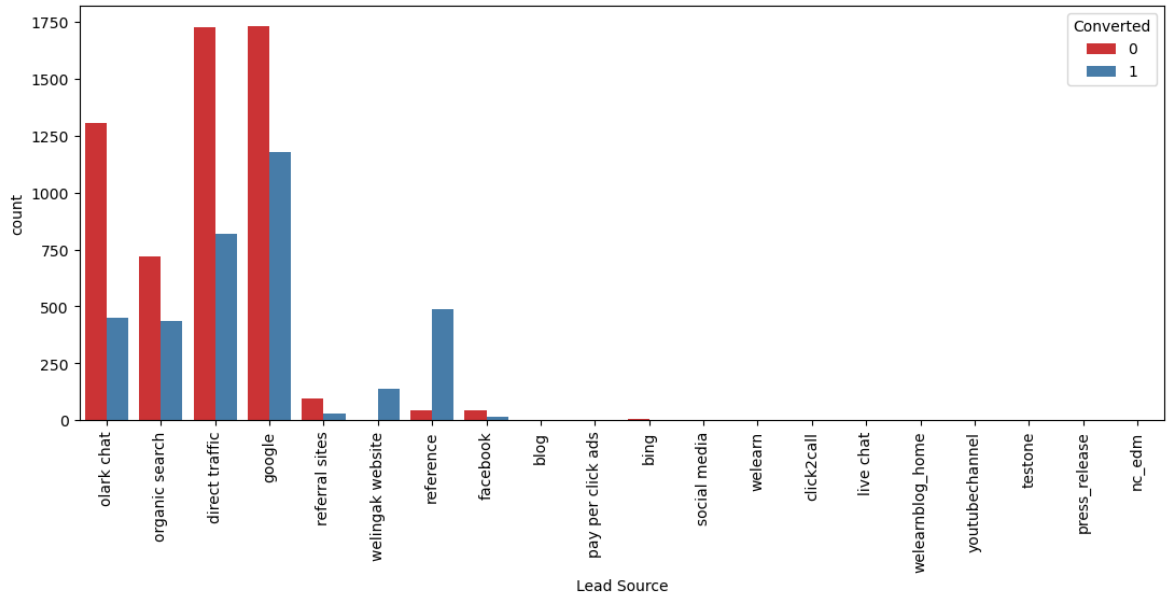
DATA CLEANING AND MANIPULATION

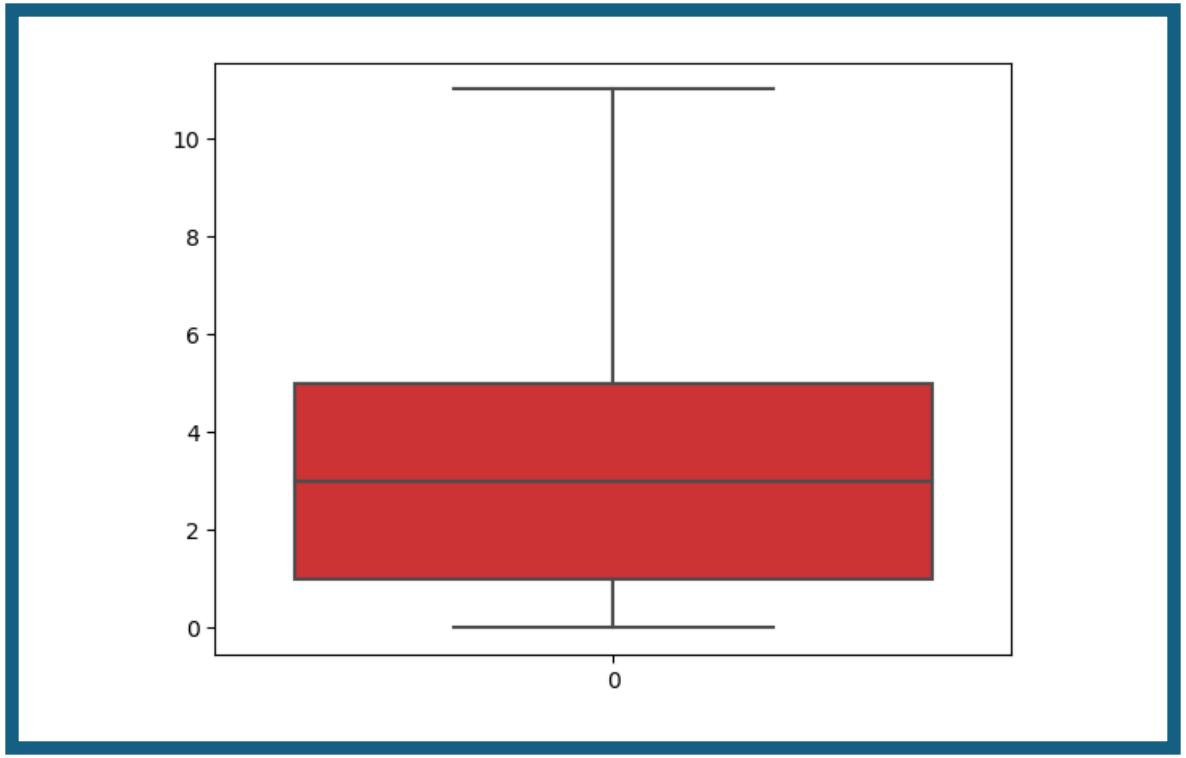
- Total Number of Rows =37, Total Number of Columns =9240.
- Handeled missing values, Removed duplicates, Handle Outliers
- Dropping the columns having more than 40% as missing value.
- Impute missing values for numerical columns using median.
- Impute missing values for categorical columns using mode.
- Replaced 'SELECT' with NAN

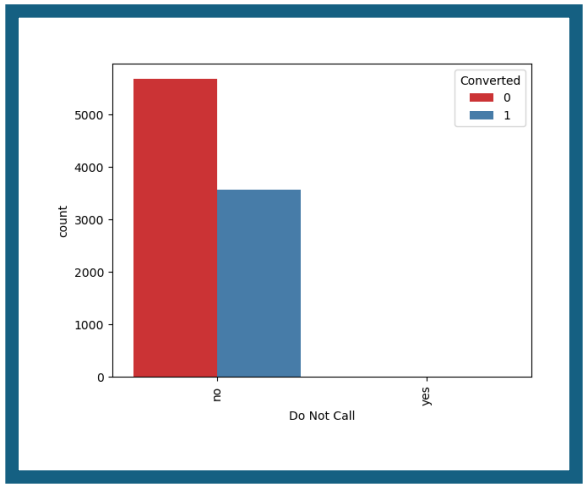
EDA (Exploratory Data Analysis)

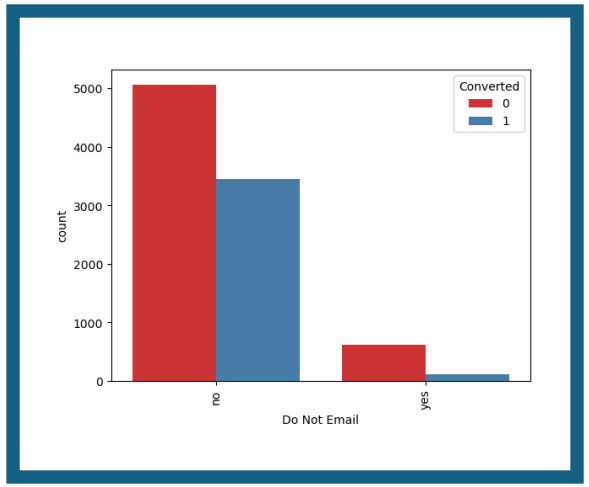
Univariate and Bivariate Analysis

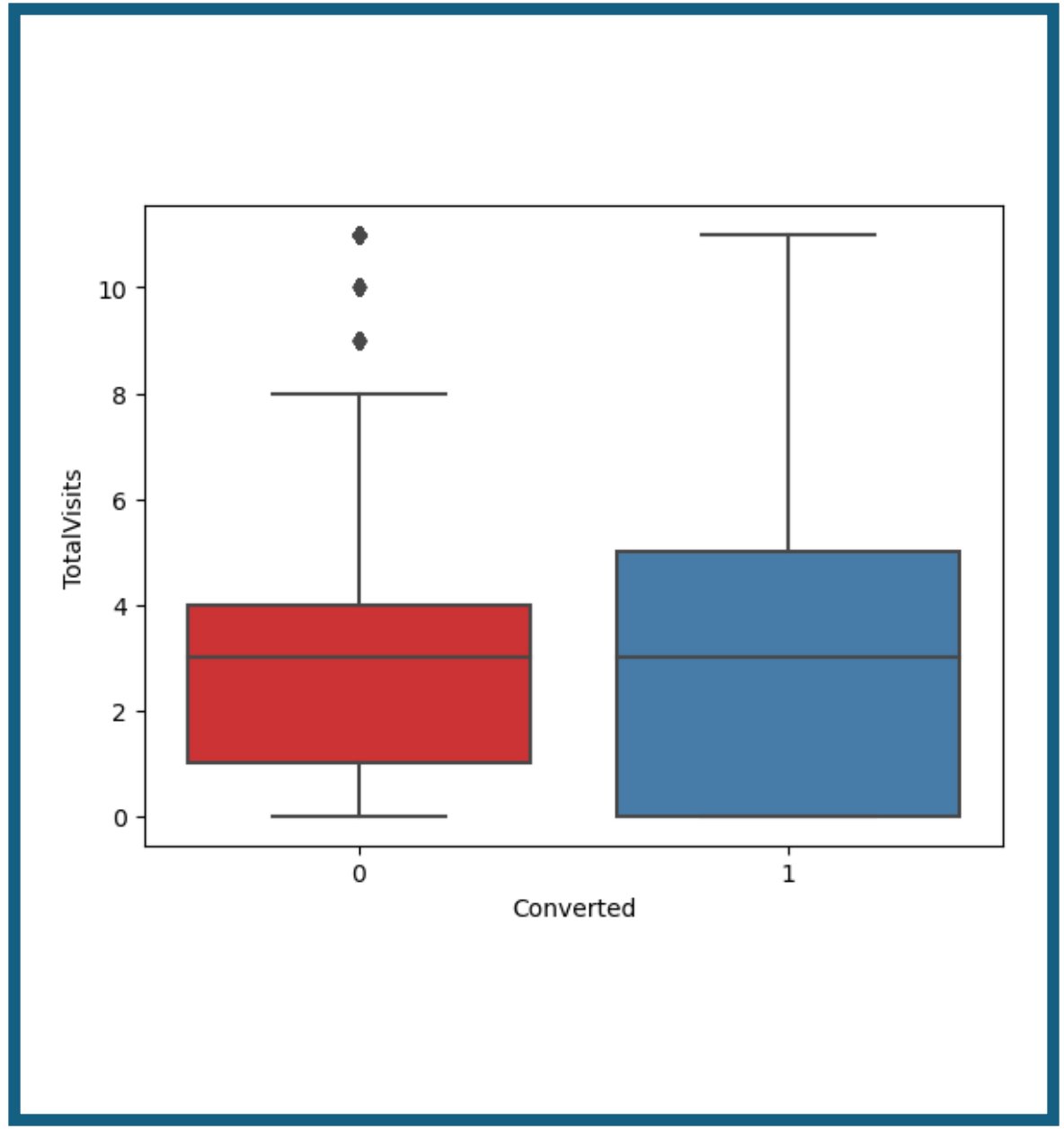


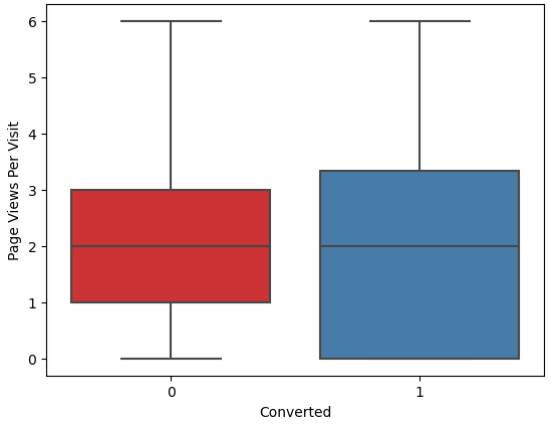


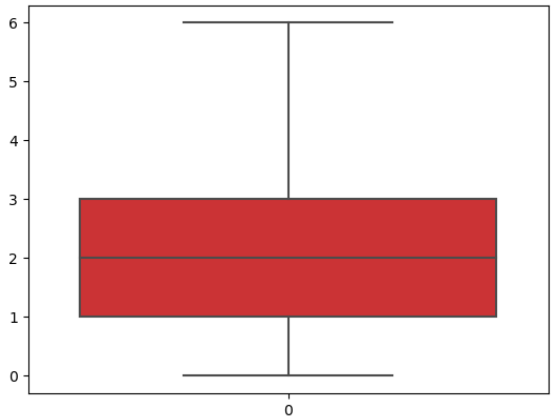

10
8
6
4
2
0
0

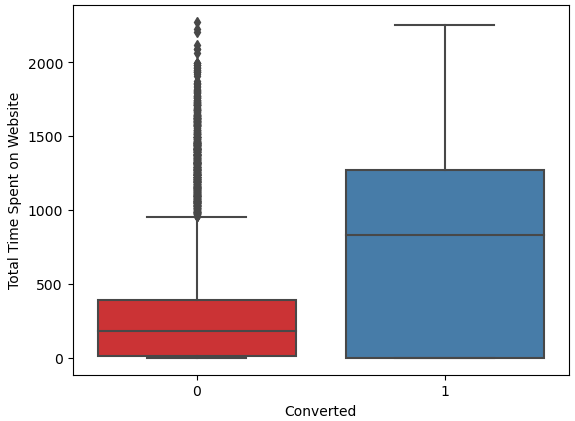

Converted
0
1
5000
4000
3000
2000
1000
0
count
no
yes
Do Not Call

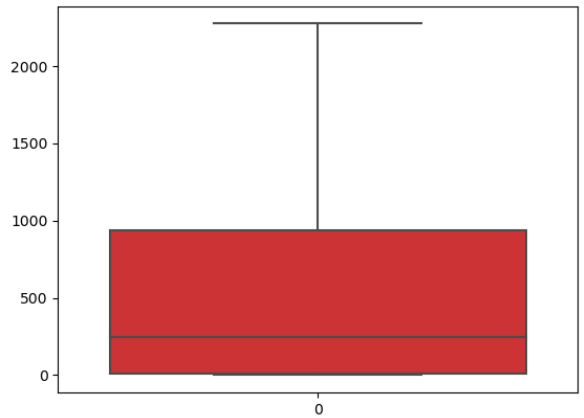

Converted
0
1
5000
4000
3000
2000
1000
0
count
no
yes
Do Not Email

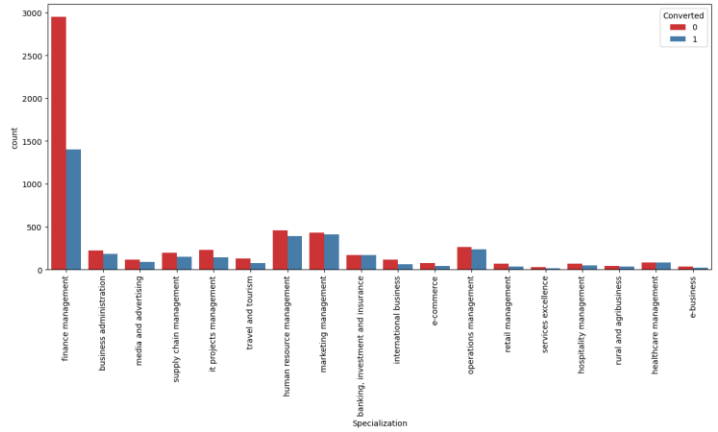

TotalVisits
10
8
6
4
2
0
0
1
Converted

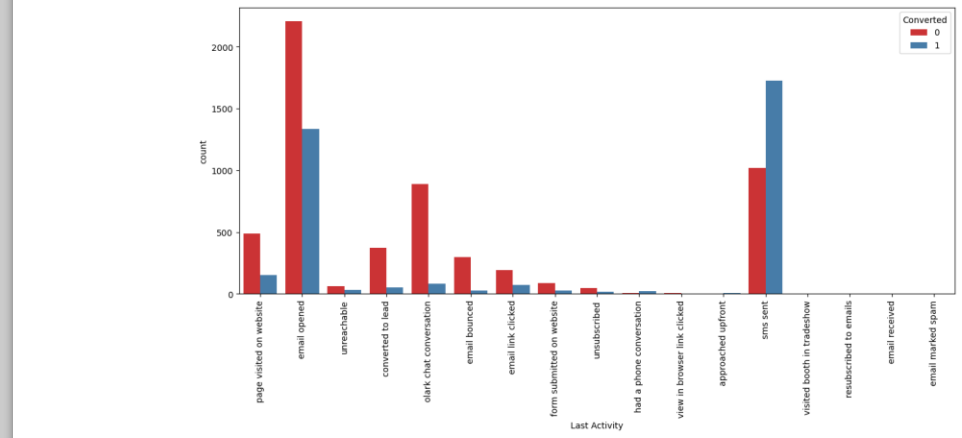

Page Views Per Visit
0
1
2
3
4
5
6
0
1
Converted

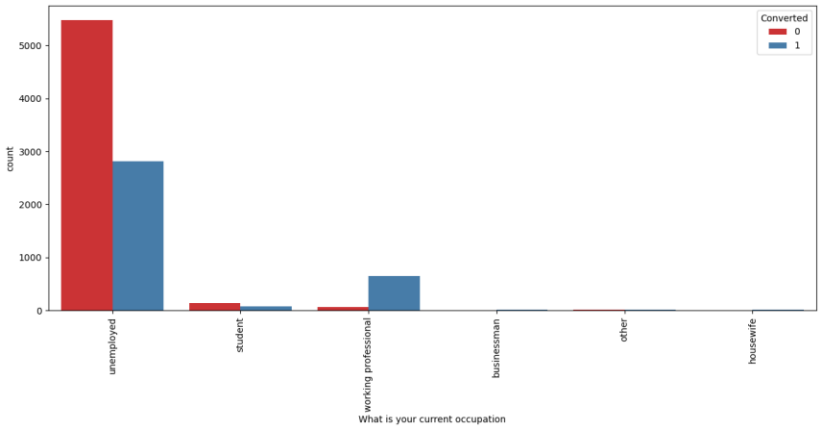

0
1
2
3
4
5
6
0

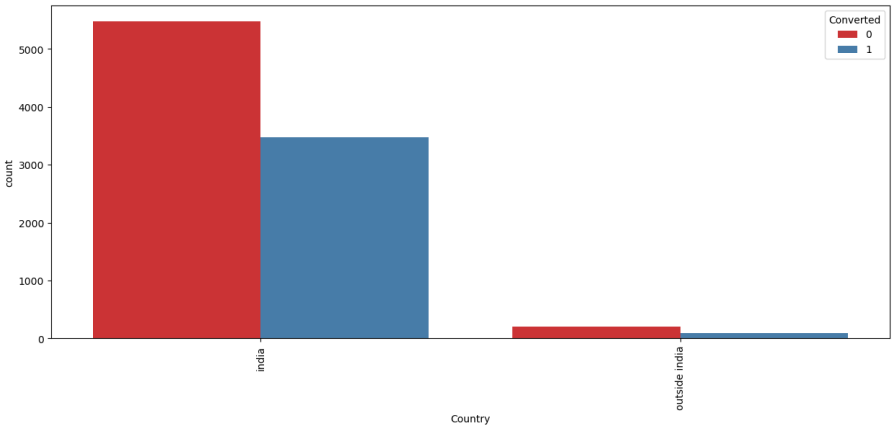

Total Time Spent on Website
0
500
1000
1500
2000
0
1
Converted

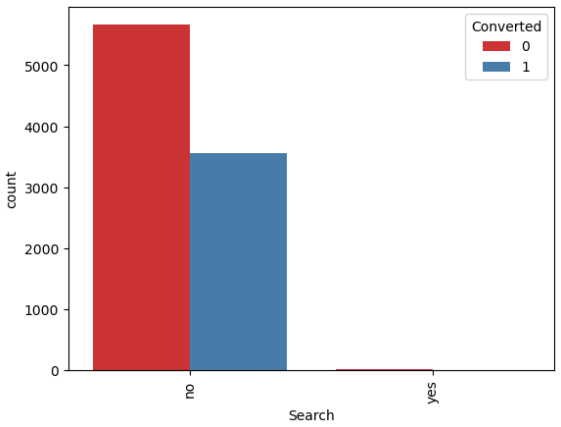

0
500
1000
1500
2000
0

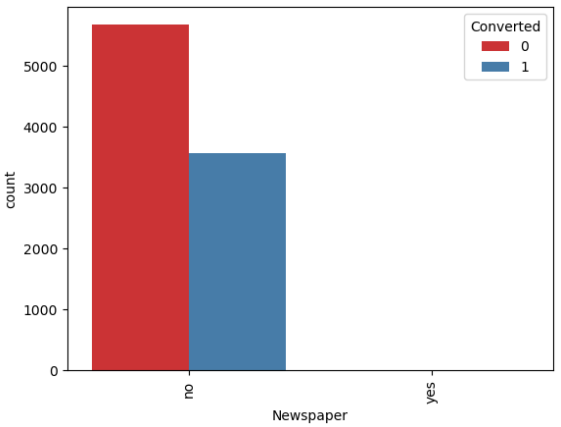

Converted
0
1
count
Specialization
finance management
business administration
media and advertising
supply chain management
it projects management
travel and tourism
human resource management
marketing management
banking, investment and insurance
international business
e-commerce
operations management
retail management
services excellence
hospitality management
rural and agribusiness
healthcare management
e-business

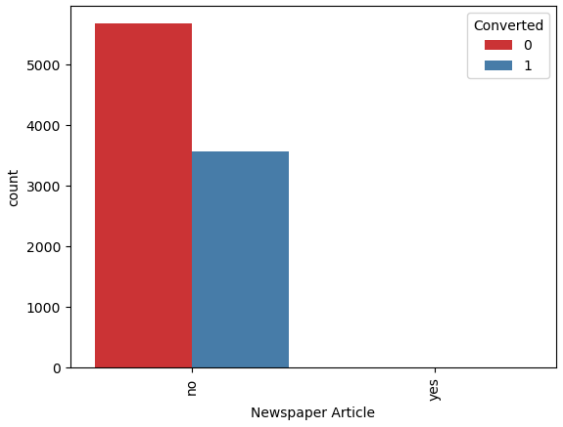

Converted
0
1
count
Last Activity
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
approached upfront
sms sent
visited booth in tradeshow
resubscribed to emails
email received
email marked spam

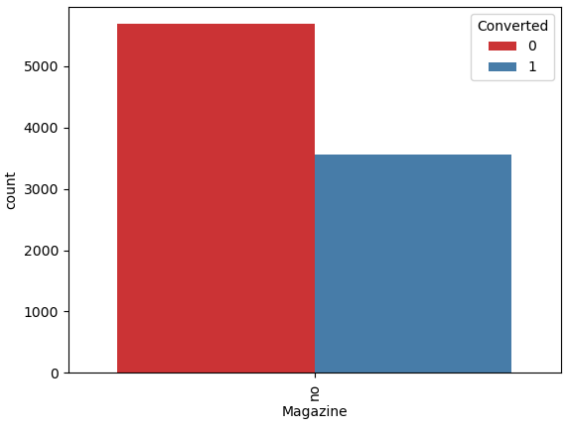

Converted
0
1
count
What is your current occupation
unemployed
student
working professional
businessman
other
housewife

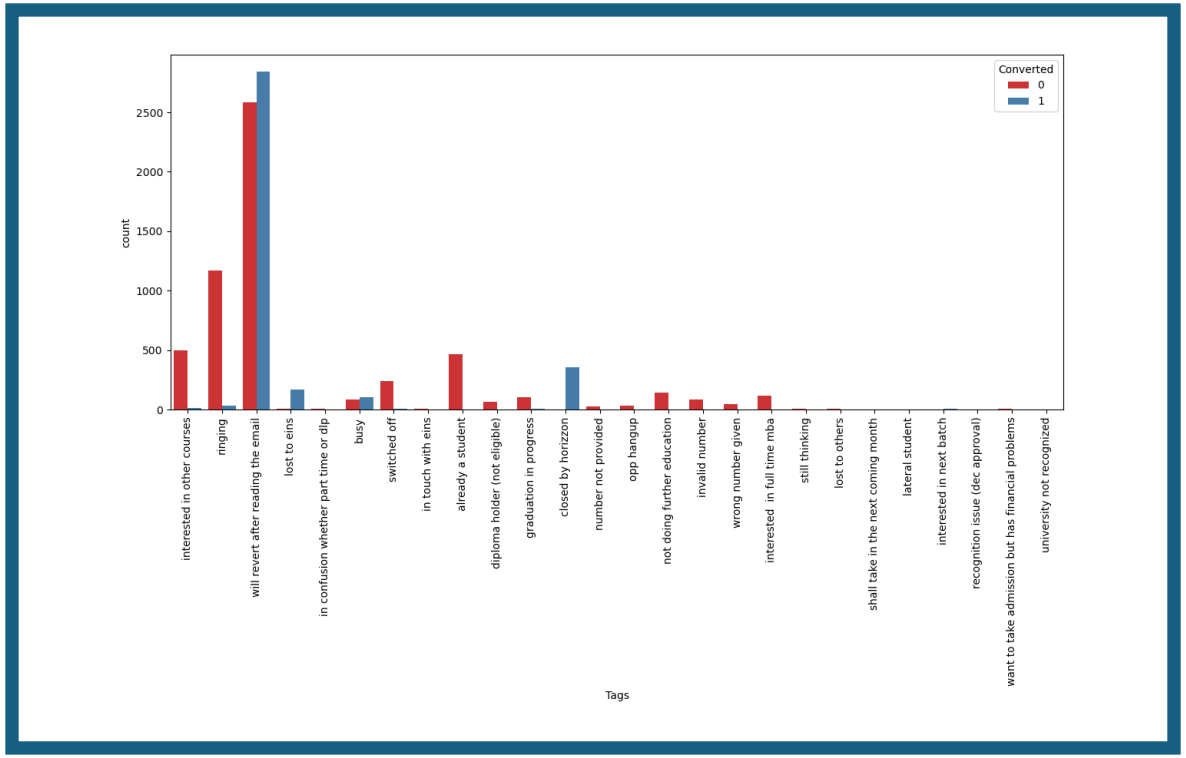

Converted
0
1
count
Country
india
outside india

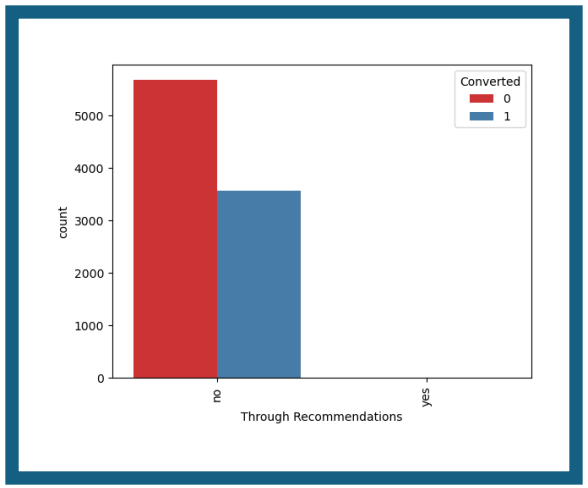

Converted
0
1
5000
4000
3000
2000
1000
0
count
no
yes
Search

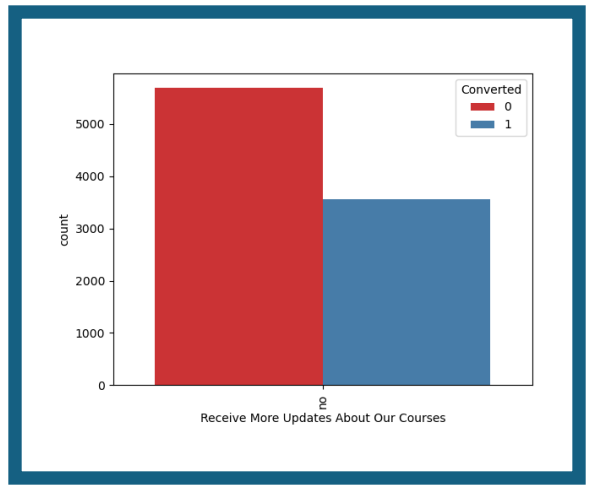

Converted
0
1
5000
4000
3000
2000
1000
0
count
no
yes
Newspaper

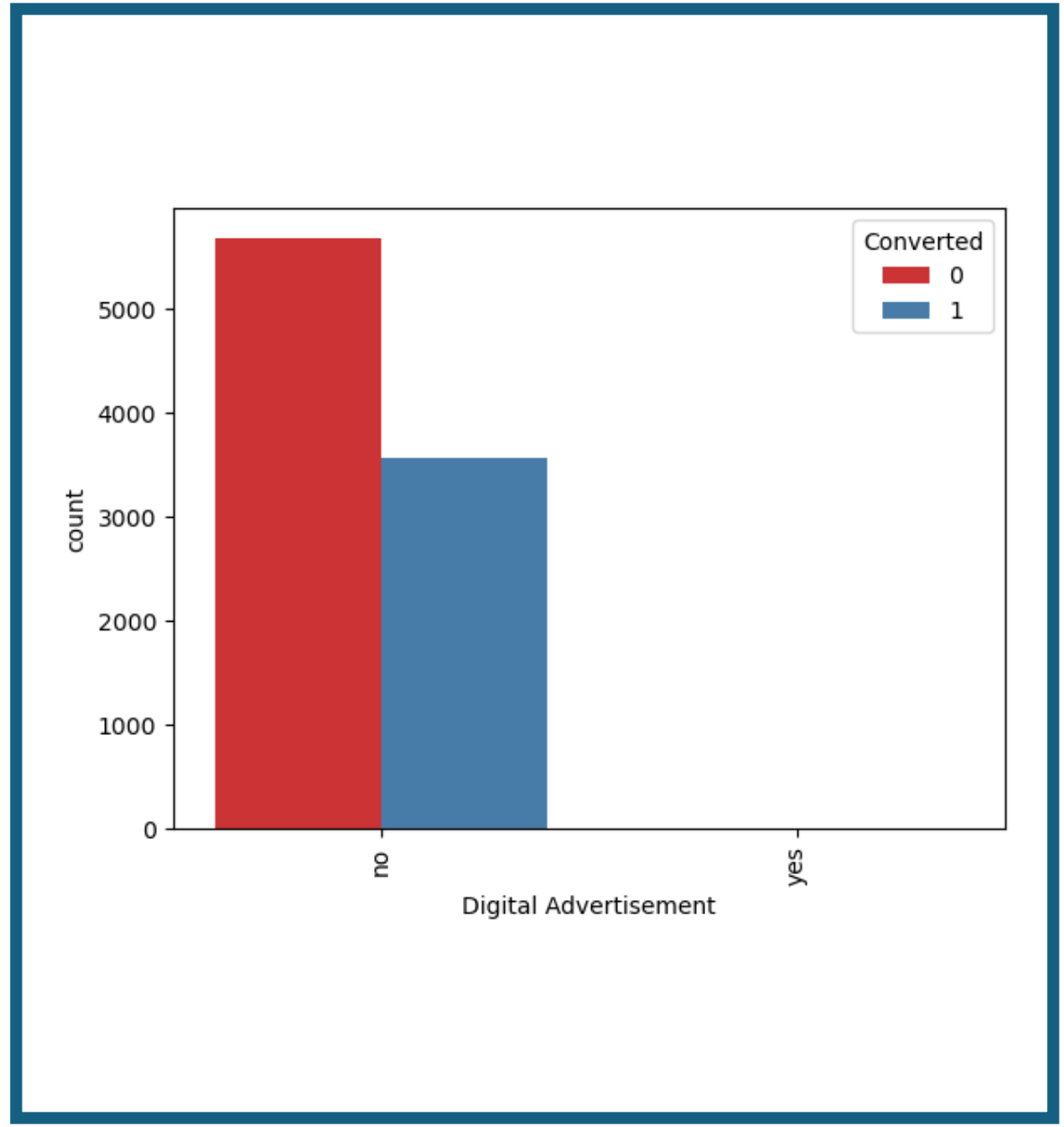

Converted
0
1
5000
4000
3000
2000
1000
0
count
no
yes
Newspaper Article

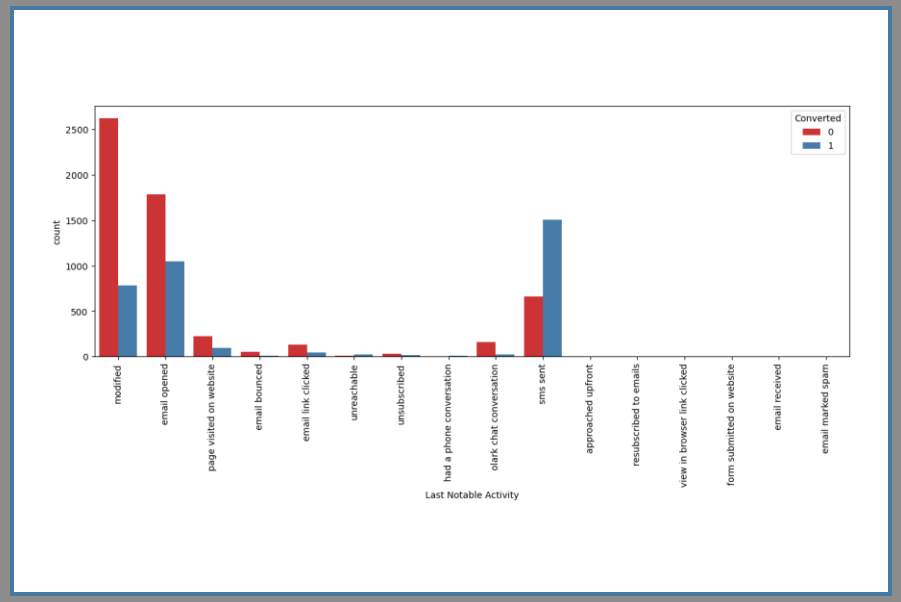

Converted
0
1
5000
4000
3000
2000
1000
0
count
no
Magazine

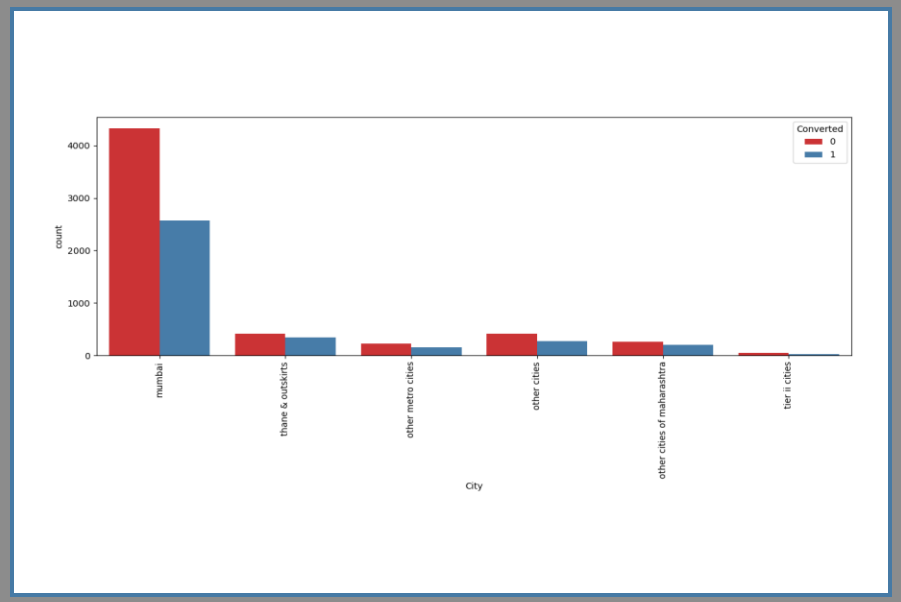

Converted
0
1
count
2500
2000
1500
1000
500
0
interested in other courses
ringing
will revert after reading the email
lost to eins
in confusion whether part time or dlp
busy
switched off
in touch with eins
already a student
diploma holder (not eligible)
graduation in progress
closed by horizzon
number not provided
opp hangup
not doing further education
invalid number
wrong number given
interested in full time mba
still thinking
lost to others
shall take in the next coming month
lateral student
interested in next batch
recognition issue (dec approval)
want to take admission but has financial problems
university not recognized
Tags

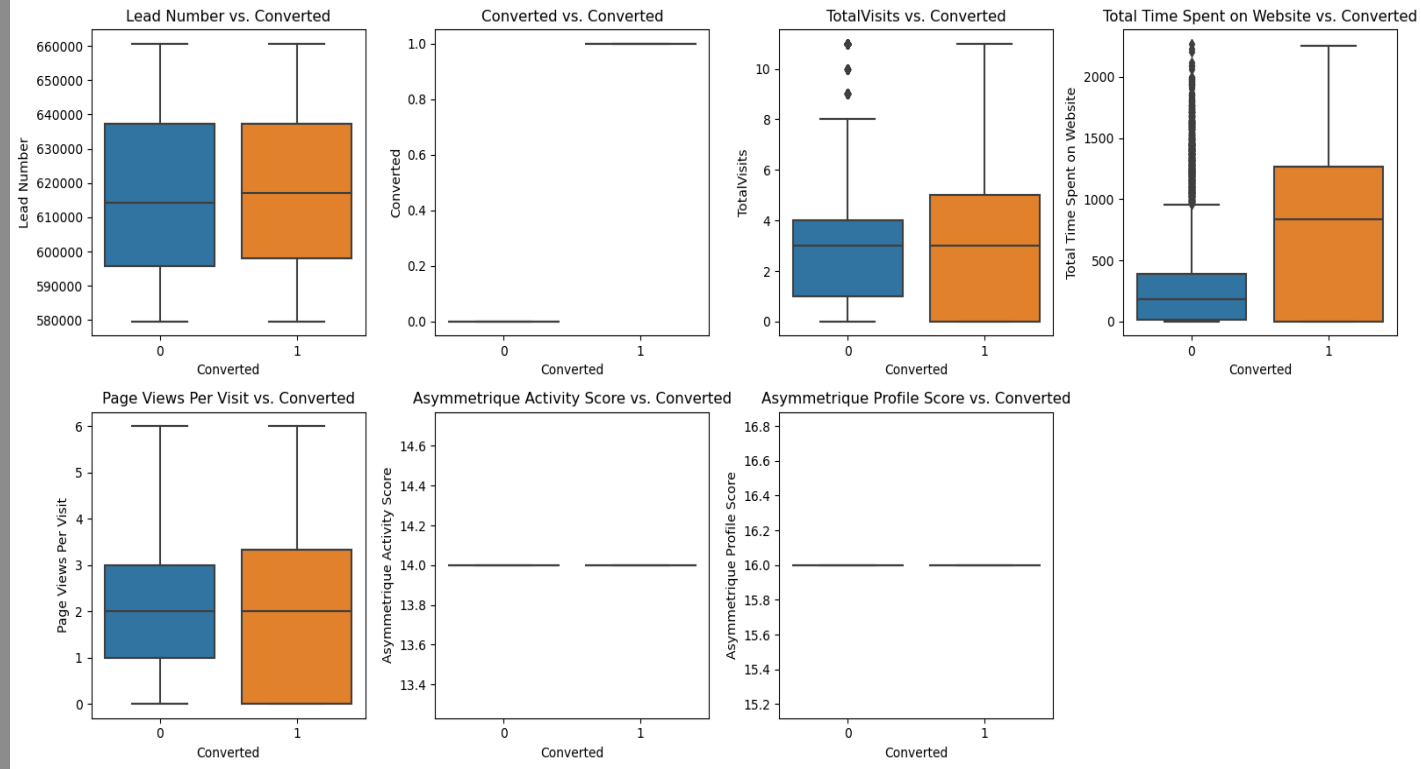

Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Through Recommendations


Converted
0
1
count
5000
4000
3000
2000
1000
0
no
Receive More Updates About Our Courses


Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Digital Advertisement


Converted
0
1
count
2500
2000
1500
1000
500
0
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received
email marked spam
Last Notable Activity


Converted
0
1
count
4000
3000
2000
1000
0
mumbai
thane & outskirts
other metro cities
other cities
other cities of maharashtra
tier ii cities
City


Lead Number vs. Converted
Lead Number
660000
650000
640000
630000
620000
610000
600000
590000
580000
0
1
Converted
Converted vs. Converted
Converted
1.0
0.8
0.6
0.4
0.2
0.0
0
1
Converted
TotalVisits vs. Converted
TotalVisits
10
8
6
4
2
0
0
1
Converted
Total Time Spent on Website vs. Converted
Total Time Spent on Website
2000
1500
1000
500
0
0
1
Converted
Page Views Per Visit vs. Converted
Page Views Per Visit
6
5
4
3
2
1
0
0
1
Converted
Asymmetrique Activity Score vs. Converted
Asymmetrique Activity Score
14.6
14.4
14.2
14.0
13.8
13.6
13.4
0
1
Converted
Asymmetrique Profile Score vs. Converted
Asymmetrique Profile Score
16.8
16.6
16.4
16.2
16.0
15.8
15.6
15.4
15.2
0
1
Converted

Multivariate Analysis

INFERENCE

Total Visits: Converted vs. Non-Converted: The scatter plots and density plots indicate that both converted and non-converted leads have a wide range of total visits. However, there seems to be a slight concentration of converted leads with higher total visits, suggesting that leads who visit the website more frequently have a higher chance of conversion.

Total Time Spent on Website: Engagement Indicator: Converted leads tend to spend more time on the website. The density plot shows a higher concentration of converted leads in the upper range of total time spent, indicating that time spent on the website is a strong indicator of conversion likelihood.

Page Views Per Visit: Exploration Behavior: Converted leads generally have a higher number of page views per visit compared to non-converted leads. This suggests that leads who explore more pages during their visits are more engaged and have a higher probability of converting.

Asymmetrique Activity Score and Asymmetrique Profile Score: Constant Values: Since these scores are constants (14 for Activity Score and 16 for Profile Score), they do not provide any meaningful variance or insight when comparing converted and non-converted leads. Therefore, they do not contribute to distinguishing between the two groups.

HYPOTHESIS TESTING

Summary

Hypothesis 1: Total Time Spent on Website

- p-value: 6.061532215246037e-285
- Conclusion**:** Reject the null hypothesis
- Implication**:** Significant difference in total time spent on the website between leads that converted and those that did not.

Hypothesis 2: Lead source and conversion

- p-value: 2.2610765566902964e-212
- Conclusion: Reject the null hypothesis
- Implication: Significant association between lead source and lead conversion.

Hypothesis 3: Total Visits

- p-value: 8.731613124367009e-06
- Conclusion: Reject the null hypothesis
- Implication: Significant difference in the number of total visits between leads that converted and those that did not.

Hypothesis 4: Page view per visit

- p-value: 0.6112393152811724
- Conclusion: Fail to reject the null hypothesis
- Implication: No significant difference in page views per visit between leads that converted and those that did not.

Hypothesis 5: City and conversion

- p-value: 8.784900567282934e-05
- Conclusion: Reject the null hypothesis
- Implication: Significant association between the city of the lead and lead conversion.

Hypothesis 6: Asymmetrique Activity Score

- p-value: nan
- Conclusion: Fail to reject the null hypothesis
- Implication: No significant difference in the Asymmetrique Activity Score between leads that converted and those that did not.

Key Findings:

- **Significant Factors:** Time spent, Lead source, Total visits, City.
- **Non-significant Factors:** Page views per visit, Asymmetrique Activity Score.

DATA CONVERSION

Numerical Variables are Normalised

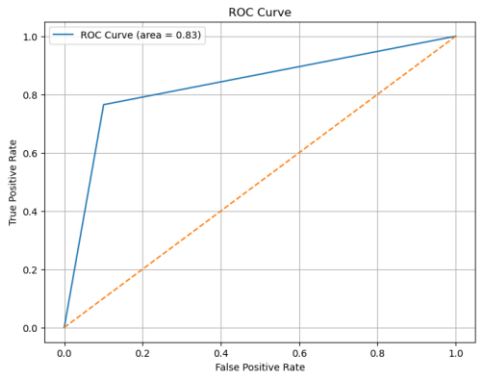
Dummy Variables are created for object type variables

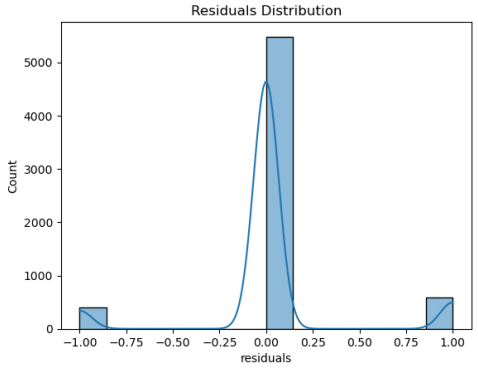
Total Rows for Analysis: 9240

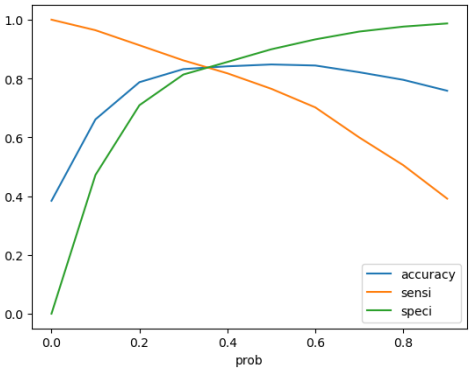
Total Columns for Analysis: 96

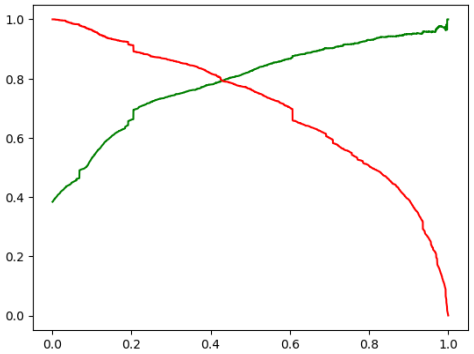
MODEL BULDING

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 20 variables as output
- Building Model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5
- Build 4 Models and choosed the best among that 4 Models based upon the business needs.
- Predictions on test data set
- Overall accuracy 83.6%









ROC Curve

Finding Optimal Cut off Point

- First graph shows the ROC curve with ROC = 83%.
- Second graph shows the concentration of residuals around zero which shows model has a good fit, with predictions being close to actual values.
- Optimal cut off probability is that probability where we get balanced sensitivity and specificity. From the third graph it is visible that the optimal cut off is at 0.35.
- Fourth graph shows the trade-off curve between precision and recall.

RESULT

Comparing Value obtained from Train and Test set

Train Data:

Accuracy : 83.7 %

Sensitivity : 84.8 %

Specificity : 82.9 %

Test Data:

Accuracy : 83.6 %

Sensitivity : 86.0 %

Specificity : 82.1 %

Leads which should be contacted or focused upon

The customers which should be contacted are the customers whose "Lead Score" is equal to or greater than 85. They can be termed as 'Hot Leads'.

There are 573 leads which can be contacted and have a higher chance of getting converted

CONCLUSION

- **It was found that the variables that mattered the most in the potential buyers are (In descending order) with respect to highest coefficients (which indicate the strongest positive impact on the target variable) are:**

Lead Origin_lead add form 3.241959
What is your current occupation_working professional 2.326409
Total Time Spent on Website 1.048084
Lead Source_olark chat 0.994948
Last Activity_converted to lead -1.107568
Specialization_hospitality management -1.115232
Last Activity_olark chat conversation -1.173468
Do Not Email -1.547412
Last Notable Activity_page visited on website -1.708281
Asymmetrique Activity Index_03.low -1.784327
Last Notable Activity_email opened -1.791193
Last Notable Activity_modified -1.873820
Last Notable Activity_email link clicked -1.924711
Last Notable Activity_olark chat conversation -1.996727
Lead Quality_might be -2.648833
Lead Quality_not sure -3.385105
Lead Quality_worst -5.183895

RECOMMENDATIONS

- The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
- The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.
- The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.
- The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.
- The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.
- The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
- The company should not make calls to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.
- The company should not make calls to the leads whose Specialization was "Others" as they are not likely to get converted.
- The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.